Speaker trait characterization in web videos: Uniting speech, language, and facial features

Weninger, Felix; Wagner, Claudia; Wöllmer, Martin; Schuller, Björn; Morency, Louis-Philipp

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

SPEAKER TRAIT CHARACTERIZATION IN WEB VIDEOS: UNITING SPEECH, LANGUAGE, AND FACIAL FEATURES

Felix Weninger[1], *Claudia Wagner*[2], *Martin Wöllmer*[1,3], *Björn Schuller*[1,2], *and Louis-Philippe Morency*[4]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Institute of Information Systems, JOANNEUM RESEARCH, Graz, Austria
[3] BMW Group, Munich, Germany
[4] Institute for Creative Technology, University of Southern California, USA

ABSTRACT

We present a multi-modal approach to speaker characterization using acoustic, visual and linguistic features. Full realism is provided by evaluation on a database of real-life web videos and automatic feature extraction including face and eye detection, and automatic speech recognition. Different segmentations are evaluated for the audio and video streams, and the statistical relevance of Linguistic Inquiry and Word Count (LIWC) features is confirmed. In the result, late multi-modal fusion delivers 73, 92 and 73 % average recall in binary age, gender and race classification on unseen test subjects, outperforming the best single modalities for age and race.

Index Terms— speaker classification, computational paralinguistics, multi-modal fusion

1. INTRODUCTION

In the emerging field of computational paralinguistics, speaker characterization according to a variety of traits (such as age, gender, height, personality, spoken language, nativeness, dialect, etc.) has received increasing attention [1–6], cf. [7] for an overview. Applications are manifold and include category-based retrieval in large audio archives, as well as adaptation of commercial voice portal systems and spoken language dialogue systems to provide customized experience to specific target groups [7]. Furthermore, recognition of gender from faces has a long tradition in computer vision [8]. Many approaches deal with static images (cf. [8] for an overview), but recently attention has shifted to videos [9–11]. Age and ethnicity / biological race have received less attention, but first studies report resonable performance in lab conditions [12] and also in real-life surveillance videos [10]. Finally, a significant body of literature exists on text categorization according to attributes of the writer, for example, in weblogs [13, 14]. However, this kind of research is usually limited to natural language processing; in order to exploit the promising correlations found in these studies for *spoken* language processing, the effects of automatic speech recognition have to be considered. A notable exception is [15] who use the output of an automatic speech recognizer to determine demographic traits.

In summary, most of the work done so far is uni-modal; we only know of a few studies performing multi-modal fusion for gender recognition such as [16, 17]. To our knowledge, multi-modal age or race recognition have not been attempted before. Hence, in this study we propose the joint extraction of audio, video and linguistic features from the audio and video streams, and multi-modal fusion, to provide a holistic way of recognizing traits of the speaker in web videos. By that, this paper is also a first attempt at a comparative evaluation of modalities in the challenging task of real-life speaker characterization. Full realism is provided by using a multi-modal database of videos obtained from social media platforms (cf. Section 2) and automatic extraction of synchronized audio and video descriptors, as well as linguistic features by means of ASR (cf. Section 3). The experimental setup and the results are discussed in Section 4 before concluding in Section 5.

The authors would like to thank Kenji Sagae, Congkai Sun and Simone Hantke for their valuable contributions. Correspondence should be addressed to weninger@tum.de.

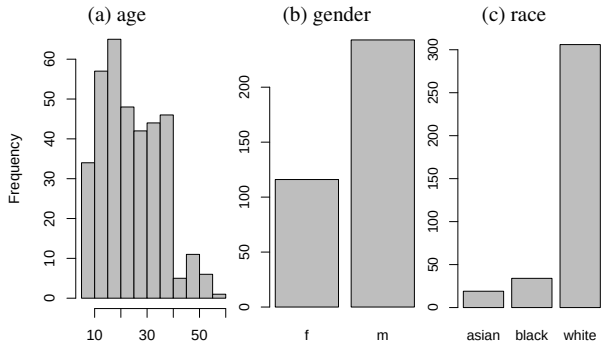
2. THE ICT-MMMO CORPUS

The ICT-MMMO (Multi-Modal Movie Opinion) corpus[1] was introduced in [18] and was originally collected for opinion mining in movie reviews. The dataset contains 345 multimodal videos where one person is speaking directly at the camera, expressing their opinion and/or stating facts related to a specific movie. Videos were collected from the social media websites YouTube and ExpoTV and contain a large number of movie reviews authored by non-professionals. Furthermore, all video clips are manually transcribed to extract spoken words as well as the start time of each spoken utterance. We use the ICT-MMMO database because (a) it is a realistic data set with the diversity, multi-modality and ambient noise characterizing real-world speaker classification, and (b) the availability of a manual transcription allows evaluating 'ground truth' linguistic features without ASR accuracy as a confounding factor, next to performance in a fully automatic setting. Speakers are from different ethnic backgrounds, yet all express themselves in English. The length of the videos varies from 1-3 minutes. All videos are converted to MPEG-4 format with a frame rate of 30 Hz for the video channel, and PCM with 16 kHz sampling rate for the audio channel.

For the purpose of this study, the speaker attributes in the videos were annotated by a 27 year old female with a background in audio-visual pattern recognition. The annotated attributes include subject ID, race ('white' = Caucasian / white Hispanic, 'black' = mostly African-American, and Asian), age (in 5 year intervals) and gender (male / female). Note that we aim at classifying biological race, not ethnicity which is a much more complex phenomenon. The annotator was instructed to use acoustic, linguistic and visual cues in her decision. A total of 275 unique subjects are identified in the 345 videos. The distribution of the annotated attributes among the videos is shown in Figure 1.

[1] The full database will be made available online with video links and transcriptions: `http://multicomp.ict.usc.edu/`

Fig. 1: Distribution of age, gender and race labels in the ICT-MMMO corpus.



3. MULTI-MODAL FEATURE EXTRACTION

3.1. Audio Features

From the audio channel, the 1 582-dimensional feature set of the INTERSPEECH 2010 Paralinguistic Challenge [7] is extracted. This set has been designed for a variety of speaker classification tasks, especially age and gender classification. It will subsequently be denoted by 'IS10'.

Features are obtained by extracting low-level descriptors (LLDs) at 100 frames per second using window sizes from 25 ms to 60 ms, then applying segment-wise functionals intended to capture time variation in a single feature vector that is independent of the segment length. In this study, segments correspond to speaker turns (utterances between speech pauses of more than 0.5 s) or alternatively, the entire audio track of the video. In the ongoing, these units of analysis will be referred to as 'turn level' or 'episode level'.

Low-level descriptors include spectral features, cepstral features (Mel-frequency cepstral coefficients, MFCCs, describing timbre of the voice), prosodic features including loudness and fundamental frequency (F0) and finally voice quality features, including jitter and shimmer, to characterize the 'roughness' of the voice. The LLDs are smoothed by moving average low-pass filtering with a window length of three frames, and their first order regression coefficients are added following the formula from [19]. This 'brute-force' combination of LLDs and functionals yields 16 zero information features which are discarded, e. g., minimum F0 (always zero). The computed LLDs and applied functionals are summarized in Table 1. We use our open-source feature extractor openSMILE [20] that also provided the features for the above named Challenge.

3.2. Video Features

For video feature extraction, each frame is transformed to an 8-bit gray scale image. Then, we perform frame-by-frame face detection using a cascade of local binary pattern (LBP)-based classifiers as implemented in OpenCV[2]. If a face is detected, we compute the centers of the eyes using a cascade of hair wavelet classifiers as implemented in openCV. The default classifier parameters are used. If two eyes have been detected, we rotate the face such as to align the eyes to the horizontal axis, and crop the face region to a square with a width of 1.6 times the distance between the eyes. In case the eye detection fails, we simply crop the image to the the detected face region. Then, the cropped is resized to a 60x60 image using bilinear interpolation.

[2] `http://opencv.willowgarage.com`

Table 1: The official 1 582-dimensional acoustic feature set ('IS10') of the INTERSPEECH 2010 Paralinguistic Challenge [7]: 38 low-level descriptors with regression coefficients, 21 functionals. Abbreviations: DDP: difference of difference of periods, LSP: line spectral pairs, Q/A: quadratic, absolute.

Descriptors	Functionals
PCM loudness	max. / min. (position)
MFCC [0–14]	arith. mean, std. deviation
log Mel Freq. Band [0–7]	skewness, kurtosis
LSP [0–7]	lin. regression slope, offset
F0 by Sub-Harmonic Sum	lin. regression error Q/A
F0 Envelope	quartile 1 / 2 / 3
Voicing Probability	quartile range 2–1 / 3–2 / 3–1
Jitter local	percentile 1 / 99 ($\approx$ min. / max.)
Jitter DDP	percentile range 99–1
Shimmer local	up-level time 75 / 90

The cropped and resized face images are transformed to 8-bit LBP images. Each pixel in the LBP image corresponds to a binary word that is computed using the $\mathrm{LBP}_{8,1}$ and $\mathrm{LBP}^{u2}_{8,1}$ operators. For a pixel with gray scale value g_c, the $\mathrm{LBP}_{P,R}$ operator computes a P-bit binary word by means of

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} u(g_p - g_c) 2^p$$

where $u(\cdot)$ is the heavyside function and g_p are the gray scale values of P equally spaced pixels surrounding the center pixel on a circle of radius R, in a defined order. The $\mathrm{LBP}^{u2}_{P,R}$ operator is based on $\mathrm{LBP}_{P,R}$ and maps all 'non-uniform' binary words to a single binary word, where 'non-uniform' means that the word has more than two bit transitions ($0 \rightarrow 1$ or $1 \rightarrow 0$), indicating a non-uniform texture around the pixel.

The above procedure transforms a video into a sequence of LBP images, from which histograms are computed. For comparability with the audio feature extraction (cf. above), 'turn level' histograms are computed across the frames within speaker turns, and 'episode level' histograms are computed from all frames of the video. The histograms have one bin per binary word, thus 256 in case of the $\mathrm{LBP}_{8,1}$ image and 59 for $\mathrm{LBP}^{u2}_{8,1}$. We compute the histograms only from the eye-aligned face images; in case that there are no frames with detected eyes we resort to the histogram of all face images. For the sake of readability, the corresponding histogram features will be denoted by LBP and LBP-U, respectively.

3.3. Linguistic Features

In this study we focus on Linguistic Inquiry and Word Count (LIWC) features. Previous research has established that physical and psychological functioning are associated with the content of writing [21]. In order to analyze such content in an objective and quantifiable manner, Pennebaker and colleagues developed a computer based text analysis program, known as LIWC [22]. LIWC uses a word count strategy searching for over 2300 words within any given text. The words have previously been categorized by independent judges into over 70 linguistic dimensions. These dimensions include standard language categories (e. g., articles, prepositions, pronouns including first person singular, first person plural, etc.), psychological processes

Table 2: Relevance of LIWC features generated from manual transcription and ASR, for age, gender, and race, sorted by effect size using only statistical significant effects with $p < 0.05$.

(a) manual transcription

Rank	Age	Gender	Race
1	filler	pers. pron.	filler
2	prepositions	conjunctions	questions
3	time	social	body
4	assent	auxiliary verbs	prepositions
5	pers. pron.	filler	insight

(b) ASR

Rank	Age	Gender	Race
1	prepositions	pers. pron.	conjunctions
2	conjunctions	pronouns	insight
3	time	I	negative emotion
4	friend	filler	certain
5	article	article	causation

(e. g., positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), relativity-related words (e. g., time, verb tense, motion, space), and traditional content dimensions (e. g., sex, death, home, occupation). In this work we use those 70 linguistic dimensions[3] as features. For reference purposes, we also extracted simple bag-of-words (BoW) features. We removed English stop words, used the Snowball stemmer[4] for English, which is an improved version of the algorithm described in [23], and term frequency $\times$ inverse document frequency (TFIDF) weighting. Only episode level linguistic feature vectors are considered in this study.

To assess both the usefulness of linguistic features as such, and their performance in a real-life setting, we generate linguistic features from the manual transcription of the ICT-MMMO database, and we alternatively apply an ASR system to obtain the transcriptions automatically. The ASR system is similar to the system used in [24] and was trained on the ICT-MMMO corpus in a three-fold cross-validation scheme (cf. Section 4). The ASR system uses left-to-right Hidden Markov Models with three emitting states per phoneme and 16 Gaussian mixtures per state. We applied tied-state cross-word triphone models with shared state transition probabilities and a back-off bigram language model, all trained on the training partition of the respective fold. As input features, the ASR system processes 12 cepstral mean normalized MFCCs and logarithmic energy, along with their first and second order derivatives.

To confirm the validity of our approach for linguistic feature extraction to identify speaker age, gender or race, we conducted a Wilcoxon-Mann-Whitney-Test (for the 2 levels of gender), a Kruskall Wallis test (for the 3 levels of race) and computed Kendall's rank correlation for age. When using ASR, 35 LIWC features (out of 70) showed statistical significant differences for male and female speaker, while only 8 LIWC features showed significant differences for white, asian and black speakers. Interestingly, in the manual transcriptions only 26 LIWC features showed statistical significant differences for male and female speaker and 9 LIWC features showed significant differences for white, asian and black speakers. The correlation with age was significant for 16 LIWC features in ASR and manual transcriptions. This indicates that LIWC features are most promising for classifying speakers' gender but might also be useful for predicting their race and age.

In Table 2 we show the 'top five' LIWC features which were significantly different for the levels of each variable of interest, ranked according to their effect size. The effect size for the variable age is determined by Kendall's τ, while the effect sizes of the variables race and gender are determined by the chi squared and z statistics. It can be seen that mostly syntax-related (e. g., filler words, pronouns, etc.) features are relevant, rather than semantic ones. While we notice clear differences in the feature relevance in manual and ASR transcripts, the above observation holds for both types of transcripts.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The age and race annotation is transformed to binary labels for creating the corresponding classification tasks. In particular, the age task is to classify 'young' (25 years or younger) vs. 'old' speakers (30 years and older) while for the race task the 'black' and 'asian' speakers were unified to a single class due to data sparsity. The age threshold of 25 is chosen as the median of the age distribution over all instances, so as to obtain a balanced classification task. The ICT-MMMO corpus is sub-divided into three folds of approximately equal size in a subject-independent fashion (i. e., testing is always done on subjects not seen in training). The fold subdivision corresponds exactly to the one used for sentiment classification in [18].

As classifiers, we use Support Vector Machines (SVMs) with linear kernel function. SVMs are trained by the Sequential Minimal Optimization (SMO) algorithm implemented in Weka [25]. Due to the class imbalance of the gender and race tasks, up-sampling of instances is applied to reach uniform class distribution in training. The complexity constant C was determined from $\{0.01, 0.1, 1, 10\}$ in subject-independent cross-validation for each modality and set to $C = 1$ for the video modality and $C = 0.1$ for the linguistic and acoustic modalities. The remaining parameters were left at their defaults. Scores in $[0, 1]$ are obtained from the distances of test vectors to the SVM hyperplanes by means of logistic regression. The audio stream and video streams are either classified using turn level features, whereby scores for the individual turns are averaged to obtain an episode level audio and video score, or directly using episode level features. In case of multi-modal analysis, the unweighted average of the audio, video and/or linguistic scores is taken.

Following [4, 7], the accuracy and unweighted average recall (UAR) of the two classes are computed for each task. We use UAR as the primary evaluation measure since it was designed for the case of imbalanced test sets, where a high accuracy could be obtained by just picking the majority class, while the UAR then would be 50 % (for binary tasks).

4.2. Results and Discussion

In Table 3, we show the performance in classifying speakers per turn, by using the audio (A) and/or video (V) modalities. For both the age and the gender tasks, combining both modalities delivers best results (67.2 % and 89.7 % UAR, respectively). The performance on the age task is remarkable given the age distribution with a high number of instances around the median. For the gender task, the UAR improvement by multi-modal fusion over either modality (A+V: 89.7 % / A: 88.0 % / V: 75.3 % UAR), is significant at the 0.1 % level according to a one-sided z-test. For the age task, fusion delivers a significant gain over the performance of the audio stream itself (61.6 % UAR) yet cannot significantly outperform the visual modality (66.7 % UAR).

[3] http://www.liwc.net/descriptiontable1.php

[4] http://snowball.tartarus.org/

Table 3: Results for turn level classification: acoustic (A) and visual (V) modalities and late audio-visual fusion (A+V). The best UAR per task is highlighted.

[%]	Features	Age		Gender		Race	
		Acc.	UAR	Acc.	UAR	Acc.	UAR
A	IS10	61.9	61.6	88.8	88.0	66.2	48.3
V	LBP	67.0	66.7	76.4	75.3	72.9	**65.3**
V	LBP-U	63.4	63.0	72.5	71.0	68.1	61.5
A+V	IS10;LBP	67.8	**67.2**	91.0	**89.7**	74.2	56.8
A+V	IS10;LBP-U	66.1	65.4	90.7	89.2	70.0	52.5

Table 4: Results for episode level uni-modal classification: visual (V), acoustic (A) and linguistic (L) modalities using different feature sets. UAR: unweighted average recall. LBP: local binary patterns. BoW: bag-of-words. LIWC: Linguistic Inquiry and Word Count [22] features obtained from manual transcription or automatic speech recognition (ASR). /T: Turn level features and score averaging per episode. /E: Episode level features.

[%]	Features	Age		Gender		Race	
		Acc.	UAR	Acc.	UAR	Acc.	UAR
A	IS10/T	63.1	62.7	93.3	**93.2**	73.5	49.7
A	IS10/E	57.6	57.0	91.0	91.7	79.4	52.3
V	LBP/T	71.6	**71.0**	82.6	81.4	80.2	**70.3**
V	LBP/E	69.4	68.7	80.3	78.3	81.8	61.4
L	BoW	54.9	54.7	74.4	65.3	84.0	49.3
L	LIWC	65.4	64.6	67.2	65.3	73.0	67.1
L	—/ASR	58.7	57.6	76.7	75.4	67.7	58.4

Table 5: Results of episode level multi-modal late fusion (+) of visual (V), acoustic (A) and linguistic (L) modalities. Features: IS10 for A; LBP for V; LIWC for L. Trs: manual transcription. ASR: automatic speech recognition. UAR above the best single modality is highlighted.

[%]	Age		Gender		Race	
	Acc.	UAR	Acc.	UAR	Acc.	UAR
A+V	69.5	69.0	93.3	92.3	82.9	62.9
A+L/Trs	68.6	67.8	90.1	89.4	76.5	55.4
V+L/Trs	71.0	70.1	80.5	79.4	81.4	**73.4**
A+V+L/Trs	73.5	**72.9**	92.4	91.4	83.8	67.5
A+L/ASR	64.5	63.8	90.7	90.3	73.0	55.8
V+L/ASR	68.6	67.7	86.9	85.6	79.3	71.4
A+V+L/ASR	68.9	68.2	93.0	92.1	81.4	64.4

The LBP feature set yields significantly higher performance than the LBP-U feature set for all tasks, indicating an information gain by treating the frequencies of non-uniform LBP separately. Remarkably, the audio modality performs below chance level (48.3 % UAR) on the race task; this is rather surprising as the vocal tract parameters of speakers from different races are known to be different [26] and should be captured by the MFCC features. As a consequence, in our experiments we always found fusion performance for race recognition to be deteriorated by the audio modality.

By moving from turn level to episode level classification, the results displayed in Table 4 are obtained. Again, the best result for the age task (71.0 % UAR) is obtained by visual features (standard LBP) while acoustic features only deliver 62.7 % UAR. Regarding the linguistic modality, LIWC features are promising for age recognition in principle (64.6 % UAR when using manual transcriptions), yet cannot compete with acoustic or visual features in full realism (57.6 % UAR using ASR). For gender recognition, acoustic features deliver the best result (93.2 % UAR). Here, visual features are ranked second (81.4 % UAR). Interestingly, LIWC features deliver a remarkable performance of 75.4 % UAR on ASR transcripts—the latter is even higher than when using manual transcripts (65.3 % UAR) which can be attributed to the robust recognition of the simple words that have been found to be significant for gender recognition (cf. Table 2). BoW features are only competitive for the gender task, which is in accordance with the findings of [13]. For race, finally, visual features perform best, yielding a remarkable UAR of 73.0 % despite the highly imbalanced task. The second best result is obtained by LIWC features, but only when manual transcripts are used. BoW and acoustic features only perform at chance level. Overall, it is notable that extraction of turn level features and score averaging delivers more robust results than extracting episode level features for both audio and video, corroborating previous results on speaker classification [5] in an audio-visual setting.

In Table 5, the results of multi-modal fusion by score averaging are shown. For the age task, we observe a slight UAR increase of 1.9 % when combining all three modalities, but no gain if ASR is used. For the gender task, none of the fused modalities can outperform the acoustic stream itself; the three modalities, using ASR, yield 92.1 % UAR in comparison to 93.2 %. However, it is notable that LIWC features from ASR and the video stream together deliver 85.6 % UAR for the gender task. For the race task, an absolute UAR gain 3.1 % can be obtained by linguistic features, but only if the manual transcription is used. Note that these fusion results do not change significantly when using the modality with the maximum classifier confidence instead of score averaging.

5. CONCLUSIONS AND OUTLOOK

We have presented a first comparative study on the performance of acoustic, visual, and linguistic features for the challenging task of speaker trait classification in web videos. Overall best results were achieved by multi-modal fusion. While linguistic features derived from the ground truth transcription delivered remarkable performance, erroneous ASR has been shown to deteriorate their usefulness for the task at hand. The generally low performance of linguistic features might also be attributed to the restricted domain of movie reviews. Naturally, also in the extraction of acoustic and facial features there is large room for improvement regarding robustness (for example, by performing latent factor analysis for compensation of channel effects, and including head pose estimation [11, 18] or advanced face tracking, cf. [24]). The most crucial next step, however, will be to corroborate our findings by generation of a large scale data set, which can be done efficiently and semi-automatically using web crawling and annotation by active learning, based on the models presented in this paper.

6. RELATION TO PRIOR WORK

Many uni-modal approaches exist for classification of one or more of the person traits addressed in this paper: For instance, [9–12] use video, [13–15] use text, and [1, 3, 6] use audio only. Audio-visual approaches are presented in [16, 17], but only for the gender task. Acoustic-linguistic personality analysis in full realism is addressed by [5]. [18] introduces the ICT-MMMO database and performs multi-modal analysis, yet only for sentiment classification. Multi-modality as employed in our study has not been investigated, to the best of our knowledge.

7. REFERENCES

[1] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *Proc. of ICASSP*, Honolulu, Hawaii, 2007, pp. 1089–1092.

[2] I. Mporas and T. Ganchev, "Estimation of unknown speakers' height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

[3] B. Schuller, M. Wöllmer, F. Eyben, G. Rigoll, and D. Arsić, "Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits," in *Proceedings AES 42nd International Conference*, K. Brandenburg and M. Sandler, Eds., Ilmenau, Germany, 2011, pp. 89–97, Audio Engineering Society.

[4] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[5] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The Voice of Leadership: Models and Performances of Automatic Analysis in On-Line Speeches," *IEEE Transactions on Affective Computing*, 2012, http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2012.15.

[6] M. H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age Estimation from Telephone Speech using i-vectors," in *Proc. of INTERSPEECH*, Portland, OR, USA, 2012, no pagination.

[7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in Speech and Language – State-of-the-Art and the Challenge," *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, vol. 27, no. 1, pp. 4–39, January 2013.

[8] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Revisiting Linear Discriminant Techniques in Gender Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 858–864, 2011.

[9] A. Hadid and M. Pietikäinen, "Combining motion and appearance for gender classification from video sequences," in *Proc. 19th International Conference on Pattern Recognition (ICPR 2008)*, Tampa, FL, USA, 2008, IEEE, no pagination.

[10] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," in *Biometric Technology for Human Identification VII*, SPIE. 2010.

[11] M. Demirkus, D. Precup, J. Clark, and T. Arbel, "Soft Biometric Trait Classification from Real-world Face Videos Conditioned on Head Pose Estimation," in *Proc. IEEE Computer Society Workshop on Biometrics in association with IEEE CVPR*, 2012, pp. 130–137.

[12] A. Hadid, "Analyzing facial behavioral features from videos," in *Human Behavior Understanding*, A. A. Salah and B. Lepri, Eds., vol. 7065 of *Lecture Notes in Computer Science*, pp. 52–61. Springer Berlin Heidelberg, 2011.

[13] S. Nowson and J. Oberlander, "The identity of bloggers: Openness and gender in personal weblogs," in *In Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*, 2006.

[14] C. M. Bell, P. M. McCarthy, and D. S. McNamara, "Using LIWC and Coh-Metrix to Investigate Gender Differences in Linguistic Styles," in *Applied Natural Language Processing: Identification, Investigation and Resolution*, P. McCarthy and C. Boonthum-Denecke, Eds., pp. 545–556. IGI Global, 2012, doi:10.4018/978-1-60960-741-8.ch032.

[15] D. Gillick, "Can conversational word usage be used to predict speaker demographics?," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1381–1384.

[16] M. Liu, X. Xu, and T. S. Huang, "Audio-visual gender recognition," in *MIPPR 2007: Pattern Recognition and Computer Vision*, vol. 6788 of *SPIE*, pp. 678803–678803–5. 2007.

[17] M. Pronobis and M. Magimai-Doss, "Integrating audio and vision for robust automatic gender recognition.," Tech. Rep. Idiap-RR-73-2008, Idiap, November 2008.

[18] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "YouTube Movie Reviews: In, Cross, and Open-Domain Sentiment Analysis in an Audiovisual Context," *IEEE Intelligent Systems Magazine, Special Issue on Concept-Level Opinion and Sentiment Analysis*, 2013, to appear.

[19] S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK book version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Florence, Italy, October 2010, pp. 1459–1462, ACM.

[21] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[22] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[23] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 3, no. 14, pp. 130–137, October 1980.

[24] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, 2012.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[26] S. A. Xue and J. G. Hao, "Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry," *Journal of Voice*, vol. 20, no. 3, pp. 391–400, 2006.